Homework 1

PSTAT 5A: Spring 2023, with Ethan P. Marzban

Instructions

- Please submit your work to Gradescope by no later than **11:59pm on Tuesday, April 11**. As a reminder, late homework will not be accepted.
- Recall that you will be asked to upload a **single** PDF containing your work for *both* the programming and non-programming questions to Gradescope.
  - You can merge PDF files using either Adobe Acrobat, or using adobe's online PDF merger at this link.

Problem 1: Data Classification

Part (a): Classify each of the following variables as discrete, continuous, ordinal, or nominal. Justify your answers.

i. The number of accidents occurring between noon and 1pm at a particular traffic light, over the span of 100 days.

Solution: This is a discrete variable, since the set of possible values is $\{0, 1, 2, \cdots\}$ which has jumps.

ii. The amount of soda (in liters) consumed daily by 247 randomly selected UCSB students.

Solution: This is a continuous variable, since the set of possible values is $[0, \infty)$ which does not have jumps.

iii. The zip codes of 1,000 randomly selected households in California.

Solution: This is a nominal variable. Firstly, even though zip codes are numbers there is nothing inherently *numerical* about them; adding two zip codes together does not produce anything interpretable. Secondly, there is no natural ordering to zip codes; there is nothing "better" about the zip code 93117 than the zip code 93106.

Part (b)

i. Give an example of a continuous variable. Pick one that wasn't covered in lecture or part (a) of this problem.

Solution: Answers may vary; one example is "the amount of time a computer program takes to run".

ii. Give an example of a discrete variable. Pick one that wasn't covered in lecture or part (a) of this problem.

Solution: Answers may vary; one example is "the amount of times a student falls asleep during lecture".

iii. Give an example of a nominal variable. Pick one that wasn't covered in lecture or part (a) of this problem.

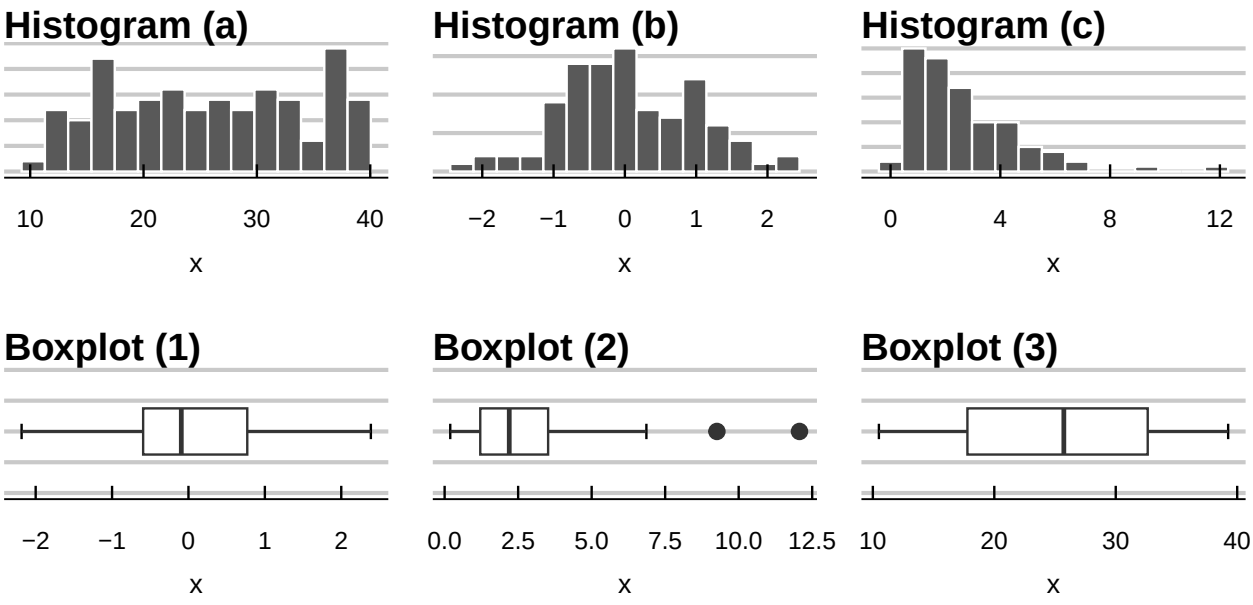
Solution: Answers may vary; one example is "favorite animal".

iv. Give an example of an ordinal variable. Pick one that wasn't covered in lecture or part (a) of this problem.

Solution: Answers may vary; one example is "highest degree (BA/BS, MA/MS, PhD, none) earned".

Problem 2: Matching

Below you will see 3 histograms, labeled (a) through (c), and 3 boxplot, labeled (1) through (3). Match each histogram to its corresponding boxplot, and justify your answers. **Hint:** In each case, think about both spread as well as how the mean compares with the median.



Solution: Here's how I thought about approaching this problem:

- Histogram (c) seems to have a large number of points that are very large compared to its median, leading us to believe there would be the presence of outliers. Only boxplot (2) has outliers, meaning we match histogram (b) with boxplot (2).

- Histogram (a) seems to have more spread than histogram (b), leading us to believe it would have a higher IQR- as such, we pick the boxplot with the widest "box", which is boxplot (3).

- This leaves Histogram (b) to be matched with boxplot (1).

There are other ways to justify this as well.

Problem 3: Appropriate Visualizations

In the parts below, you will be provided with a pair of variables x and y. For each part, identify the appropriate type or types of visualization for x, along with the appropriate type of plot to visualize the relationship between x and y.

a. x is monthly salary; y is zip code.

Solution: 'x' is numerical (continuous, specifically) meaning a histogram or boxplot would be best to visualize its distribution. 'y', as mentioned in Problem 1, is categorical (nominal), meaning to compare 'x' and 'y' we should use a side-by-side boxplot.

b. x is average commute time (in hours, including decimals); y is amount of sleep (also in hours, including decimals)

Solution: x is numerical (continuous, specifically) meaning a histogram or boxplot would be best to visualize its distribution. y is also numerical (continuous), meaning to compare x and y we should use a scatterplot.

c. x is month of the year, y is amount of ramen (in pounds) consumed.

Solution: x is categorical (ordinal), meaning a barplot would be best to visualize its distribution. y is numerical (continuous), so a side-by-side boxplot is best to visualize the relationship between x and y.

Problem 4: Associations

a. Give an example of two variables you think would have a positive association.

Solution: Answers may vary; height and weight is a common example (taller people tend to weigh more).

b. Give an example of two variables you think would have a negative association.

Solution: Answers may vary; "time since an earthquake" and "number of aftershocks per minute" is one (since aftershocks tend to occur less and less frequently as we progress farther chronologically from the main quake).

Problem 5: Transformations

The notion of **transforming data** is an incredibly important one. As an example, suppose $F = \{f_i\}_{i=1}^n$ denotes a set of temperature measurements, as recorded in Fahrenheit. If we wish to convert the measurements to Centigrade, we obtain a new set of data $C = \{c_i\}_{i=1}^n$ where each c_i is linked with a corresponding f_i through the formula

$$c_i = \frac{5}{9}(f_i - 32)$$

In general, we consider a set $X = \{x_i\}_{i=1}^n$ and a function $g : \mathbb{R} \to \mathbb{R}$, and construct a new set of data as $Y = \{y_i\}_{i=1}^n$ with $y_i = g(x_i)$. This is all we mean by a **transformation:** a function that we apply to each point in a dataset to obtain a new dataset.

a. Suppose we take a linear transformation g; i.e. we take $y_i = ax_i + b$ for some fixed constants a and b. (A concrete example of such a transformation is the conversion from Fahrenheit to Centigrade mentioned above). Show that $\overline{y} = g(\overline{x})$, using the following steps:
   i. Write down the definition of $\overline{y}$

ii. Substitute $ax_i + b$ in place of y_i
iii. Perform algebraic manipulations to obtain $\overline{y} = a\overline{x} + b = g(\overline{x})$ to complete the argument.

Solution: We follow the hints provided. The definition of $\overline{y}$ is

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}(y_1 + \cdots + y_n)$$

Since $y_i = ax_i + b$ by assumption, we have

$$\begin{aligned}
\overline{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) \\
&= \frac{1}{n}\left[(ax_1 + b) + (ax_2 + b) + \cdots + (ax_n + b)\right]
\end{aligned}$$

We can now rearrange terms:

$$\begin{aligned}
\overline{y} &= \frac{1}{n}\left[(ax_1 + b) + (ax_2 + b) + \cdots + (ax_n + b)\right] \\
&= \frac{1}{n}\left[(ax_1 + ax_2 + \cdots + ax_n) + \left(\underbrace{b + b + \cdots + b}_{n \text{ times}}\right)\right] \\
&= \frac{1}{n}\left[a(x_1 + x_2 + \cdots + x_n) + nb\right] \\
&= a \cdot \underbrace{\frac{1}{n}(x_1 + x_2 + \cdots + x_n)}_{=\overline{x}} + \frac{1}{\cancel{n}} \cdot \cancel{n}b = a\overline{x} + b
\end{aligned}$$

which completes the argument.

b. Is it always true that $\overline{y} = g(\overline{x})$? If so, provide a short proof/justification. If not, give a specific counterexample. As a hint: think about nonlinear transformations as well!

Solution: It is not true! As a simple counterexample, consider $x = \{1, 2, 3\}$ and $g(x) = x^2$, so that $Y = \{1^2, 2^2, 3^2\} = \{1, 4, 9\}$. We have $\overline{x} = 2$ and so $g(\overline{x}) = (\overline{x})^2 = (2^2) = 4$, whereas $\overline{y} = (1 + 4 + 9)/3 = 14/3 \neq 4$.

Problem 6: The Median

There is another measure of central tendency: the **median**. Here's how we compute the median of a set $X = \{x_i\}_{i=1}^n$:

i. Line up the numbers in ascending order
ii. Cross off the first and last numbers.
iii. Cross off the first and last numbers that are not crossed off.
iv. Continue until you are either left with a single number (in which case this number is the median), or we are left with a pair of numbers (in which case the median will be the mean of these two numbers).

As an example, to compute the median of the set $S = \{1, 2, 3, 3, 5, 6, 10, 11\}$: we write:

- ~~1~~, 2, 3, 3, 5, 6, 10, ~~11~~
- ~~1~~, ~~2~~, 3, 4, 5, 6, ~~10~~, ~~11~~
- ~~1~~, ~~2~~, ~~3~~, 3, 5, ~~6~~, ~~10~~, ~~11~~

So, the median is $(3 + 5)/2 = \boxed{4}$.

a. Compute the median of the set $\{1, 2, 3, 4\}$.

Solution: The median is 2.5 .

b. Compute the median of the set $\{1, 2, 3, 4, 5\}$.

Solution: The median is 3 .

c. Compute the median of the set $\{1, 2, 3, 4, 5, 6\}$.

Solution: The median is 3.5 .

d. Generalize your answers to parts (a) - (c) above to find a formula for the median of the set $\{1, 2, \cdots, n\}$, for some fixed natural number n.

Solution: The median is $\frac{n+1}{2}$.

e. It turns out (as a result from mathematics) that

$$\sum_{k=1}^{n} k = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}$$

Use this fact to compute the mean of the set $\{1, 2, \cdots, n\}$. How does this compare to the median you computed in part (d) above?

Solution: We compute

$$\overline{x} = \frac{1}{n}(1 + \cdots + n) = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$$

which is actually the same as the median!

Problem 7: Programming

Instructions

- Write your answers to this question in a new Jupyter notebook, and export your work to a PDF using the steps you saw in Lab01. **Be sure to merge this PDF with your PDF containing your work to the above questions before submitting!** (See instructions at the top of this homework).

a. Write a line of code that results in a `NameError`. Don't use the same code from Lab01!

b. Run the following code in a new cell:

```
x = 2
y = 3
x = y + 2
```

Using a comment, write down what you think `print(x)` will return after having run the above code. Then, run `print(x)` and comment on whether your initial guess was correct or not.

c. Use Google to find out where the name `Jupyter` comes from. Write your answer in a Markdown cell.

d. Create a new cell, copy the code `type(abs)`, and comment on the result. Specifically, based on the result of this code, can you identify yet another data type in addition to `float` and `int`?

e. Navigate to this (https://docs.python.org/3/library/functions.html) link, which contains a list (and description) of the functions that come built into Python (i.e. that can be used without needing to import any modules). Pick **two** functions to read up on, and write a brief description of both functions as well as an example call of each. For example:

- `abs()`: Computes the absolute value of a number. Example call:

```
abs(-103.203)
```

```
103.203
```

(Don't use `abs()` as one of the functions you choose!)

HW07 Problem 7 Solutions

PSTAT 5A, Compiled by Ethan

April 5, 2023

0.1 Part (a)

[1]:
```
# answers may vary. An example:
xyz
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[1], line 2
      1 # answers may vary. An example:
----> 2 xyz

NameError: name 'xyz' is not defined
```

0.2 Part (b)

[2]:
```
x = 2
y = 3
x = y + 2
```

[3]:
```
# after the second line, 'y' is assigned the avlue 3 meaning adding 2 to `y` is␣
↪equivalent to adding 3 to 2. As such, 'x' should be assigned the value "5".
```

[4]:
```
print(x)
```

```
5
```

0.3 Part (c)

Jupyter is a portmanteau of "Julia", "Python", and "R", which are thre three core languages that Jupyter supports.

0.4 Part (d)

[5]:
```
type(abs)
```

[5]:
```
builtin_function_or_method
```

[6]:
```
# yes, we can identify another data type: builtin_function_or_method
```

0.5 Part (e)

Answers may vary: as an example, we can look at the `round()` function, which rounds the given input to the desired number of decimals. For example,

[7]:
```
round(1.01942, 3)
```

[7]:
```
1.019
```

rounded 1.01942 to 3 decimal places.

[ ]:
```

```